BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–18

Supplementary material

Fabio Gori*, Gianluigi Folino†, Mike Jetten‡, Elena Marchiori*

† ICAR-CNR, Rende, Italy
* Radboud University Nijmegen, Dept. of Computer Science, The Netherlands
‡ Radboud University Nijmegen, Dept. of Microbiology, The Netherlands

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 DATA DESCRIPTION

Characteristics of the datasets used in our experimental analysis are given in Tables 1 and 2.

2 RESULTS

Table 3 shows the number of reads in each dataset selected using BLASTx and the total number of reads.

Tables 4, 5 and 6 report results on simulated datasets concerning taxon accuracy and sensitivity of the methods and the number of detected taxa.

Table 7 reports the number of detected taxa on real-life datasets.

Figures 1-30 contain pie charts showing the population characterization resulting from the taxonomic assignment computed by the methods. On the simulated datasets the true population distribution is also shown.

Table 1. Characteristics of the simulated data: identifier and name of the organism, size of its genome and total number of reads sampled for coverage $0.1X$. Detailed information on these datasets can be found in (Dalevi *et al.*, 2008).

M1 Organism	Genome size (bp)	Reads sampled
Clostridium phytofermentans ISDg	4,533,512	4,638
Prochlorococcus marinus NATL2A	1,842,899	1,866
Lactobacillus reuteri 100-23	2,174,299	2,371
Caldicellulosiruptor saccharolyticus DSM 8903	2,970,275	2,950
Clostridium sp. OhILAs	2,997,608	2,934
Herpetosiphon aurantiacus ATCC 23779	6,605,151	6,937
Bacillus weihenstephanensis KBAB4	5,602,503	4,158
Halothermothrix orenii H 168	2,578,146	2,698
Clostridium cellulolyticum H10	3,958,683	3,978
M2 **Organism**	**Genome size (bp)**	**Reads sampled**
Geobacter sp. FRC-32	3,982,463	4,225
Burkholderia multivorans ATCC 17616	6,979,389	7,110
Delftia acidovorans SPH-1	6,702,581	7,046
Comamonas testosteroni KF-1	5,906,374	6,189
Geobacter lovleyi SZ	3,871,860	4,300
M3 **Organism**	**Genome size (bp)**	**Reads sampled**
Shewanella putrefaciens CN-32	4,659,220	4,714
Shewanella loihica PV-4	4,602,594	4,588
Halorhodospira halophila SL1	2,678,452	2,690
Pseudomonas putida F1	5,959,964	6,407
Shewanella baltica OS195	5,310,173	5,378
Bifidobacterium longum bv. Infantis ATCC 15697	2,832,748	2,898
Stenotrophomonas maltophilia R551-3	4,544,233	4,685
Parvibaculum lavamentivorans DS-1	3,854,587	4,501

Table 2. Characteristics of real-life datasets retrieved from the metagenomics RAST server (Meyer *et al.*, 2008). The three real-life datasets containing short reads (average length of about 100bp) and are sampled using pyrosequencing on Roche 454 CS20. They have been derived from a saltern sample (Edwards *et al.*, 2006), a coral holobiont sample (Rodriguez-Brito *et al.*, 2007), and a chicken cecum sample, respectively.

Name	Saltern	Coral holobiont	Chicken cecum
Total bp	3,453,306	32,282,404	30,657,259
No. sequences	34,296	316,279	294,682
Max Seq. Length	248	269	258
Min Seq. Length	30	37	39
Average Seq. Length	100.69	102.07	104.4

Table 3. Number of reads in simulated datasets. From left to right: dataset name, the number of those reads in the dataset having at least one high-quality BLASTx alignment, as described in the paper (nr of selected reads) and the number of reads in the dataset (total nr of reads).

Dataset	Nr. of selected reads	Total nr. of reads
M1 0.1x	5,704	32,534
M1 1x	58,298	329,334
M1 4x	177,178	1,291,587
M2 0.1x	9,070	28,875
M2 1x	92,257	288,730
M2 4x	174,992	1,101,324
M3 0.1x	11,824	35,862
M3 1x	116,949	353,022
M3 4x	166,976	1,385,028
Saltern	1,675	34,296
Coral	24,941	316,279
Chicken	112,983	294,682

Table 4. Taxon sensitivity and accuracy, and number of detected taxa on M1 datasets.

M1	0.1x	1x	4x
MTR			
Phylum	100.00 33.33 (9)	100.00 15.00 (20)	100.00 10.71 (28)
Class	75.00 25.00 (12)	75.00 8.82 (34)	75.00 6.98 (43)
Order	57.14 22.22 (18)	71.43 8.77 (57)	66.67 5.26 (76)
Family	42.86 12.00 (25)	71.43 5.56 (90)	66.67 3.08 (130)
Genus	50.00 14.29 (28)	75.00 4.72 (127)	71.43 2.45 (204)
LCA			
Phylum	100.00 33.33 (9)	100.00 15.79 (19)	100.00 11.54 (26)
Class	75.00 30.00 (10)	75.00 9.38 (32)	75.00 7.50 (40)
Order	57.14 28.57 (14)	71.43 8.93 (56)	66.67 5.41 (74)
Family	42.86 15.00 (20)	71.43 5.75 (87)	66.67 3.15 (127)
Genus	50.00 16.67 (24)	75.00 5.13 (117)	71.43 2.60 (192)

Table 5. Taxon sensitivity and accuracy, and number of detected taxa on M2 datasets.

M2	0.1x	1x	4x
MTR			
Phylum	100.00 20.00 (5)	100.00 6.25 (16)	100.00 5.56 (18)
Class	100.00 22.22 (9)	100.00 8.33 (24)	100.00 8.00 (25)
Order	100.00 11.11 (18)	100.00 3.92 (51)	100.00 4.08 (49)
Family	100.00 12.00 (25)	100.00 3.95 (76)	100.00 4.11 (73)
Genus	75.00 10.34 (29)	100.00 3.42 (117)	100.00 2.59 (116)
LCA			
Phylum	100.00 20.00 (5)	100.00 6.25 (16)	100.00 5.56 (18)
Class	100.00 25.00 (8)	100.00 9.09 (22)	100.00 9.09 (22)
Order	100.00 11.76 (17)	100.00 4.00 (50)	100.00 4.17 (48)
Family	100.00 13.64 (22)	100.00 4.23 (71)	100.00 4.17 (72)
Genus	75.00 12.00 (25)	100.00 3.88 (103)	100.00 2.86 (105)

Table 6. Taxon sensitivity and accuracy, and number of detected taxa on M3 datasets.

M3	0.1x	1x	4x
MTR			
Phylum	100.00 40.00 (5)	100.00 18.18 (11)	100.00 28.57 (7)
Class	100.00 42.86 (7)	100.00 20.00 (15)	100.00 16.67 (12)
Order	100.00 31.58 (19)	100.00 16.67 (36)	100.00 7.69 (26)
Family	100.00 25.00 (24)	100.00 8.82 (68)	100.00 5.13 (39)
Genus	66.67 14.29 (28)	83.33 4.20 (119)	100.00 3.92 (51)
LCA			
Phylum	100.00 50.00 (4)	100.00 18.18 (11)	100.00 28.57 (7)
Class	100.00 50.00 (6)	100.00 21.43 (14)	100.00 16.67 (12)
Order	100.00 33.33 (18)	100.00 16.67 (36)	100.00 7.69 (26)
Gamily	100.00 27.27 (22)	100.00 9.68 (62)	100.00 5.71 (35)
Genus	66.67 21.05 (19)	83.33 4.76 (105)	100.00 5.13 (39)

Table 7. Number of detected taxa on real-life datasets

Real Life	Saltern	Coral	Chicken
MTR			
Kingdom	2	3	2
Phylum	6	17	15
Class	9	24	22
Order	12	46	32
Family	6	58	47
Genus	8	66	61
Species	15	70	133
LCA			
Kingdom	2	3	2
Phylum	4	16	15
Class	6	23	21
Order	7	40	29
Family	5	47	44
Genus	4	52	55
Species	8	58	135

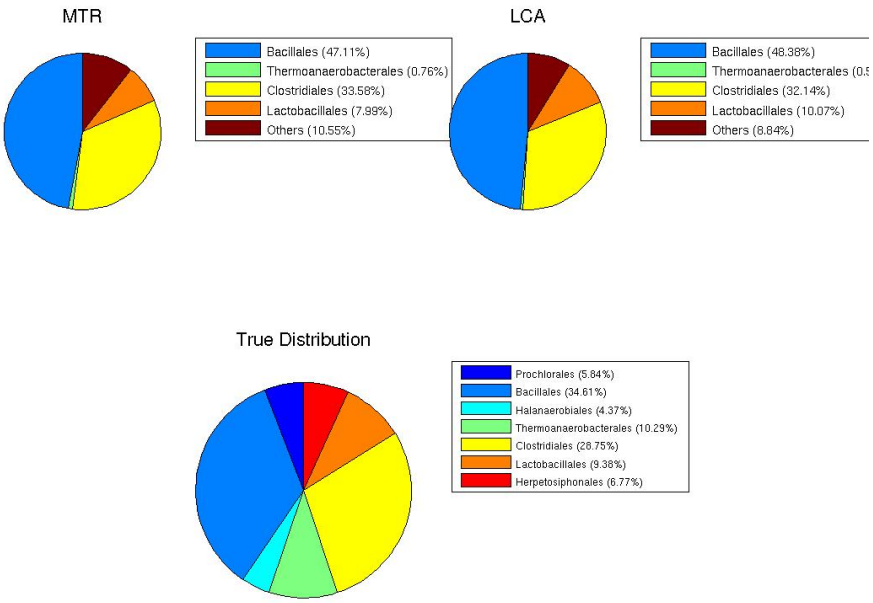

Fig. 1. Population distributions (rank Order) of M1, coverage 0.1x, by MTR and LCA, and the true population distribution.

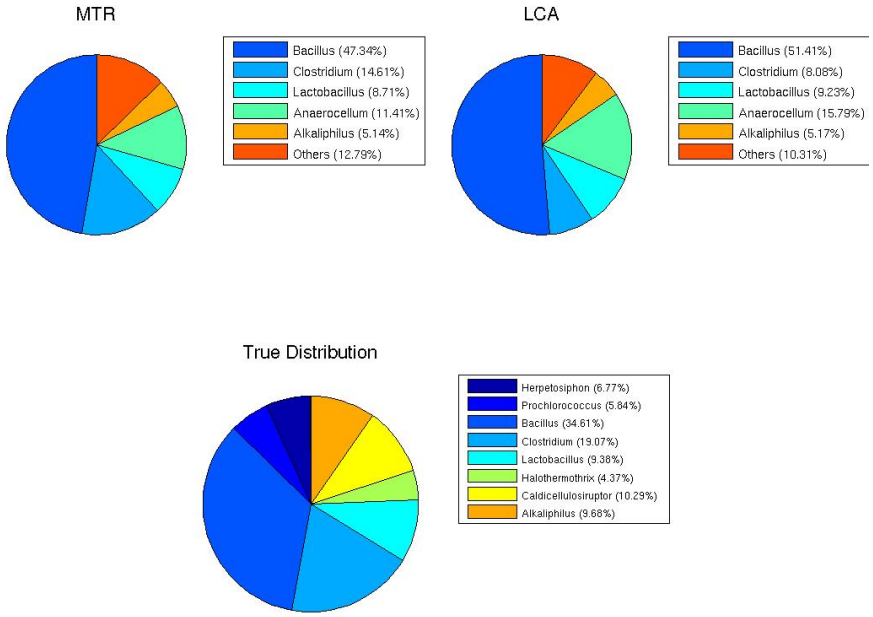

Fig. 2. Population distributions (rank Genus) of M1, coverage 0.1x, by MTR and LCA, and the true population distribution.

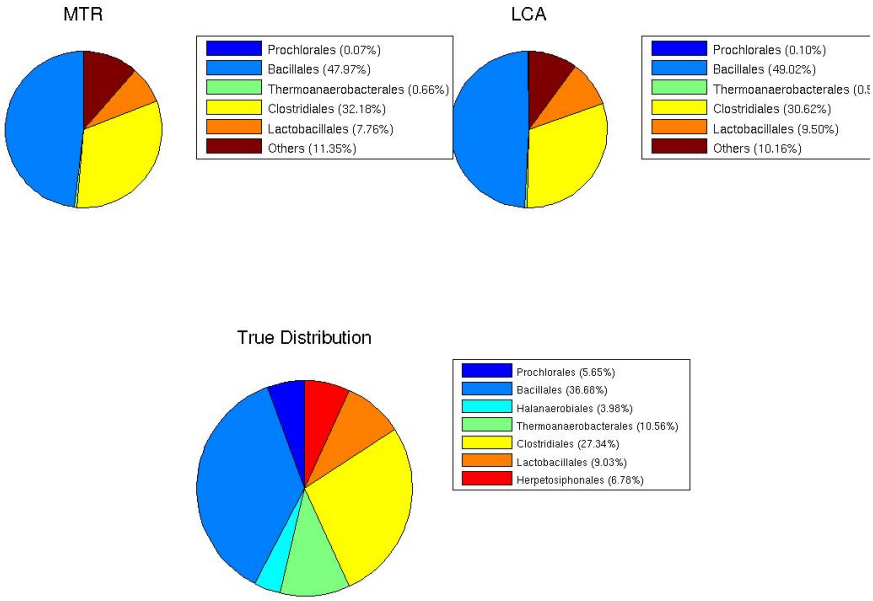

Fig. 3. Population distributions (rank Order) of M1, coverage 1x, by MTR and LCA, and the true population distribution.

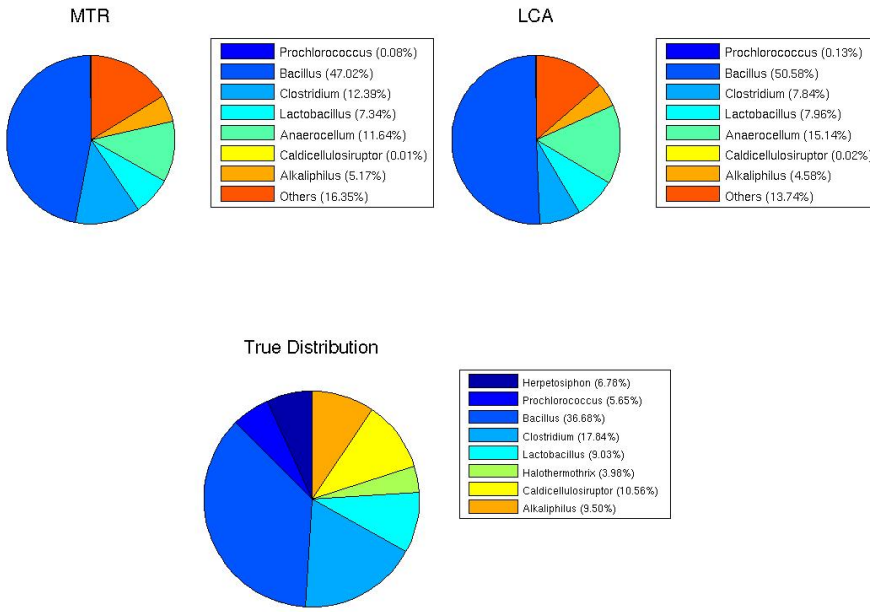

Fig. 4. Population distributions (rank Genus) of M1, coverage 1x, by MTR and LCA, and the true population distribution.

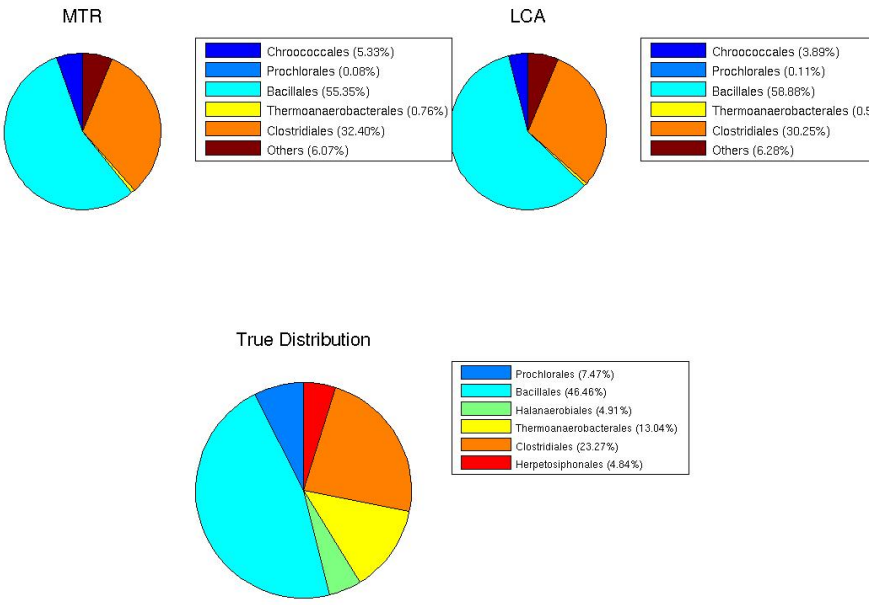

Fig. 5. Population distributions (rank Order) of M1, coverage 4x, by MTR and LCA, and the true population distribution.

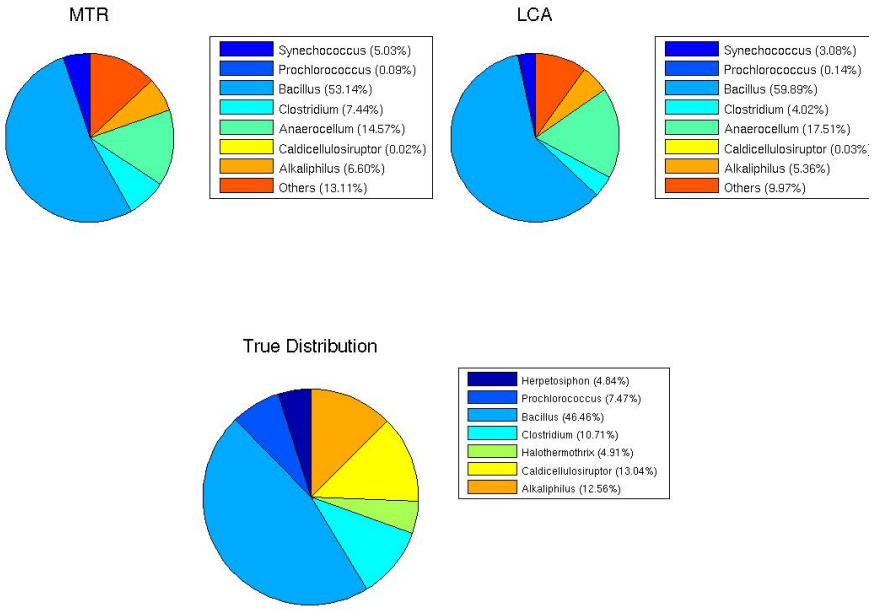

Fig. 6. Population distributions (rank Genus) of M1, coverage 4x, by MTR and LCA, and the true population distribution.

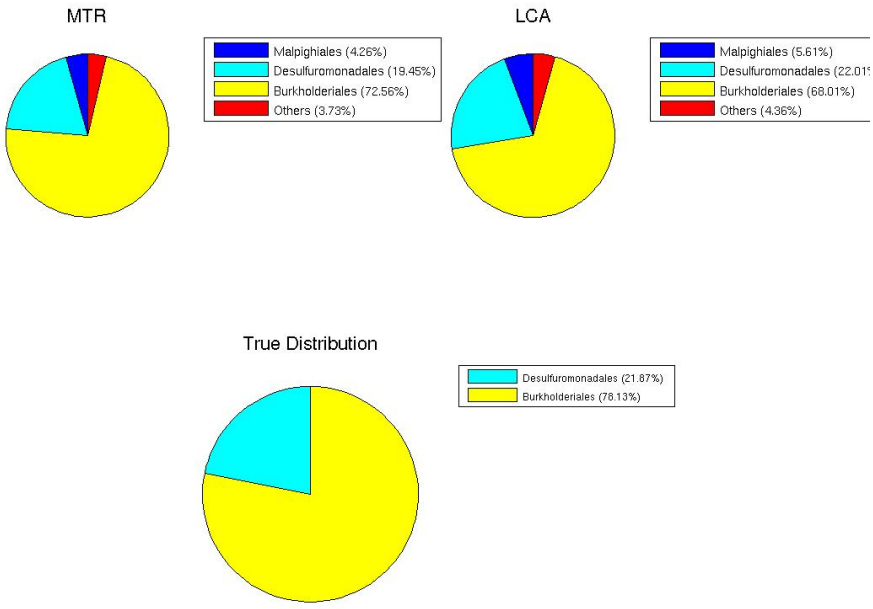

Fig. 7. Population distributions (rank Order) of M2, coverage 0.1x, by MTR and LCA, and the true population distribution.

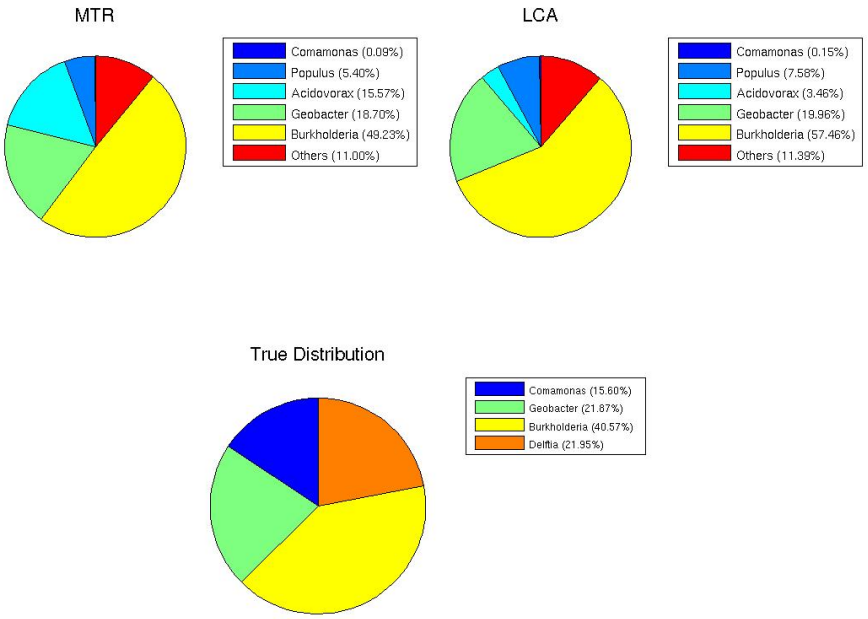

Fig. 8. Population distributions (rank Genus) of M2, coverage 0.1x, by MTR and LCA, and the true population distribution.

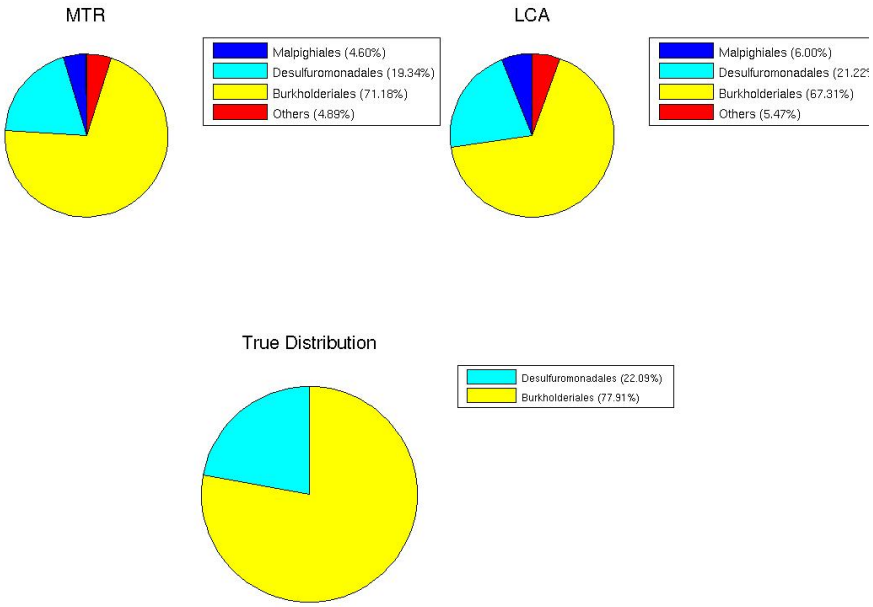

Fig. 9. Population distributions (rank Order) of M2, coverage 1x, by MTR and LCA, and the true population distribution.

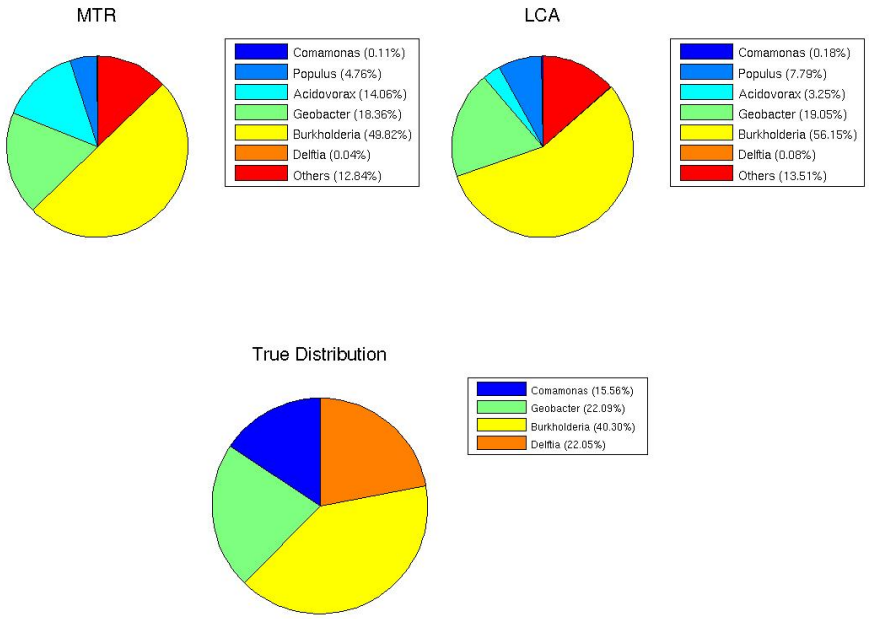

Fig. 10. Population distributions (rank Genus) of M2, coverage 1x, by MTR and LCA, and the true population distribution.

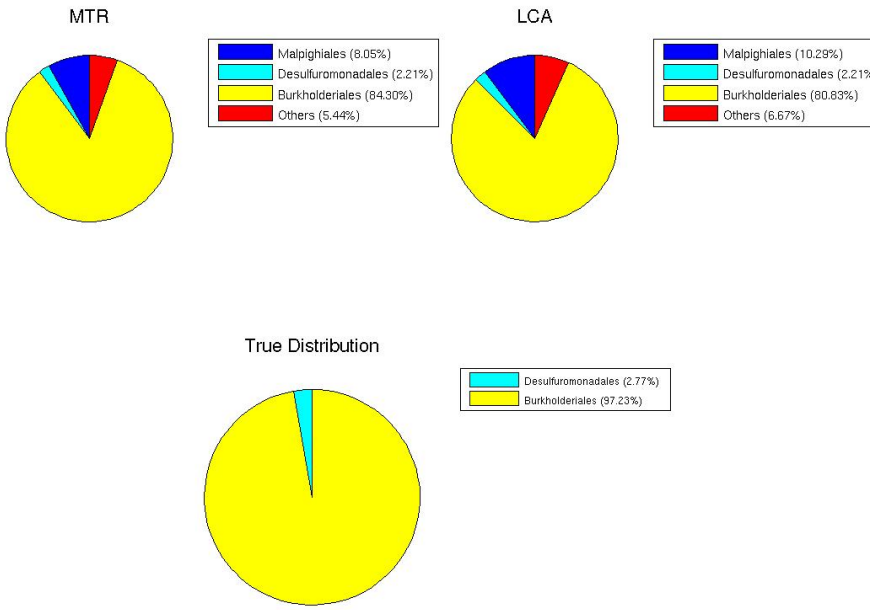

Fig. 11. Population distributions (rank Order) of M2, coverage 4x, by MTR and LCA, and the true population distribution.

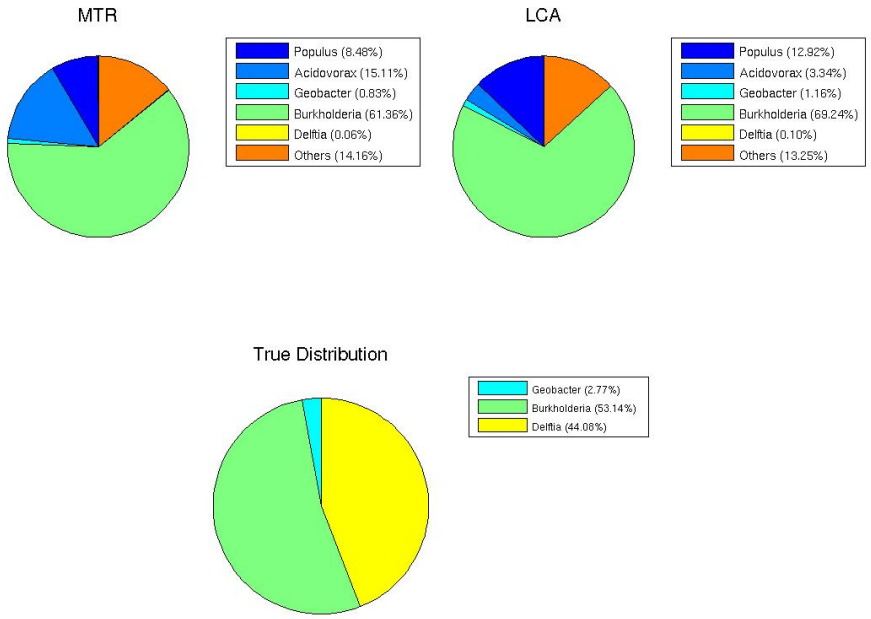

Fig. 12. Population distributions (rank Genus) of M2, coverage 4x, by MTR and LCA, and the true population distribution.

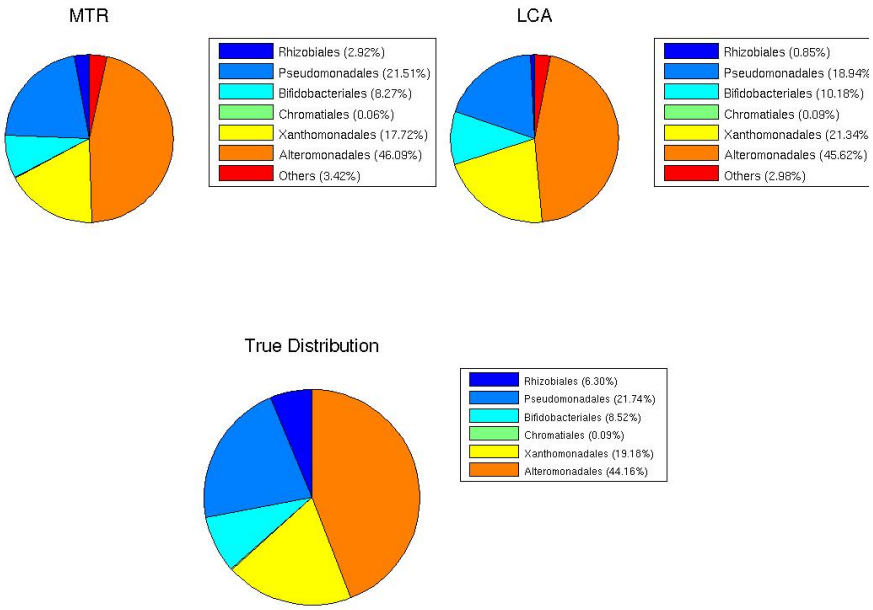

Fig. 13. Population distributions (rank Order) of M3, coverage 0.1x, by MTR and LCA, and the true population distribution.

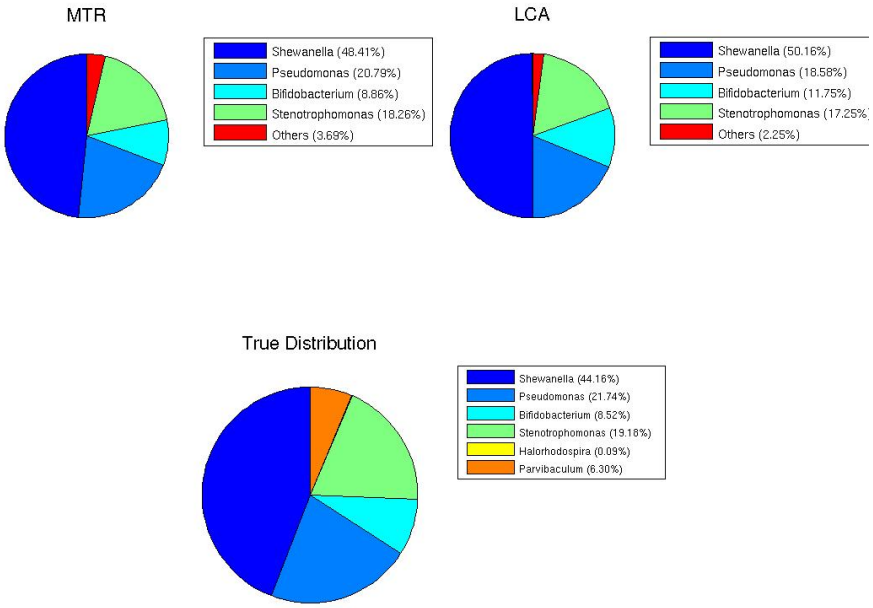

Fig. 14. Population distributions (rank Genus) of M3, coverage 0.1x, by MTR and LCA, and the true population distribution.

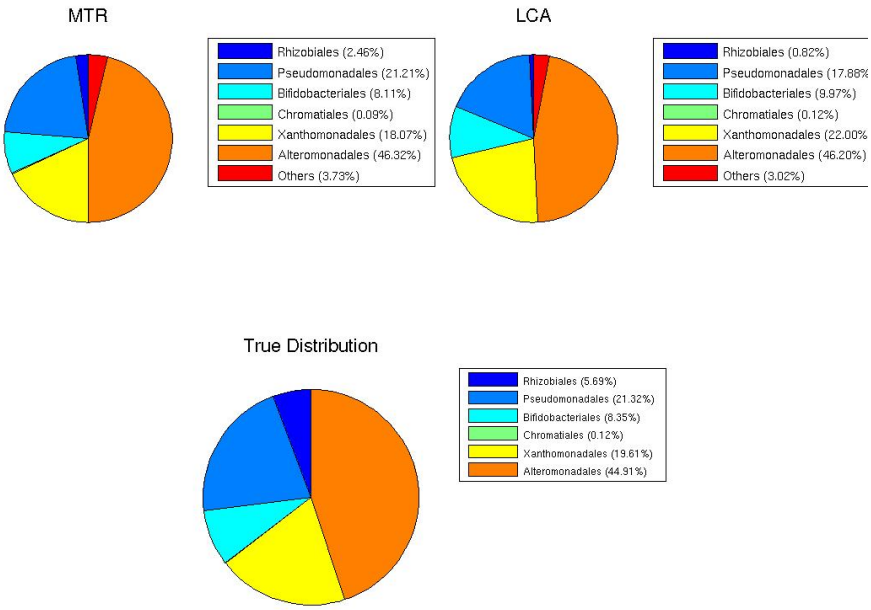

Fig. 15. Population distributions (rank Order) of M3, coverage 1x, by MTR and LCA, and the true population distribution.

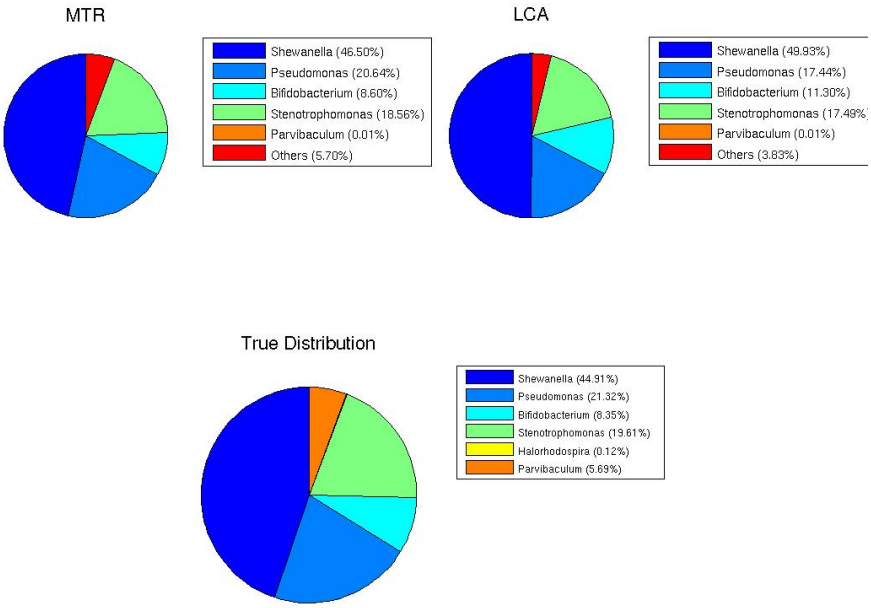

Fig. 16. Population distributions (rank Genus) of M3, coverage 1x, by MTR and LCA, and the true population distribution.

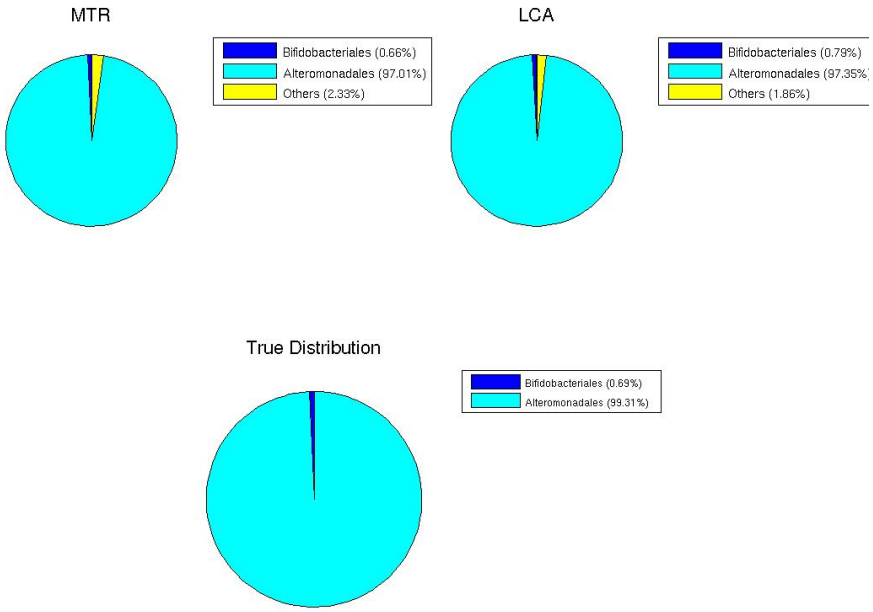

Fig. 17. Population distributions (rank Order) of M3, coverage 4x, by MTR and LCA, and the true population distribution.

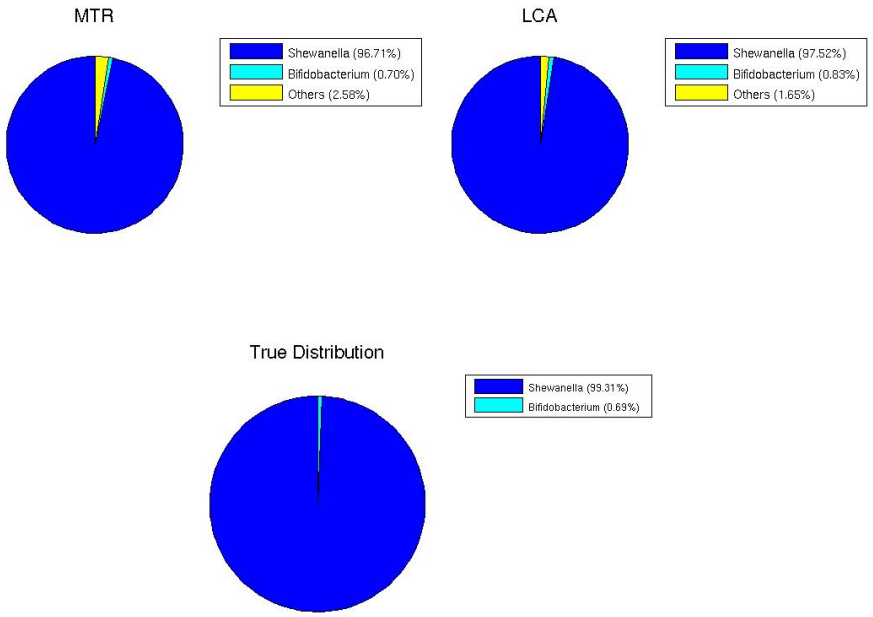

Fig. 18. Population distributions (rank Genus) of M3, coverage 4x, by MTR and LCA, and the true population distribution.

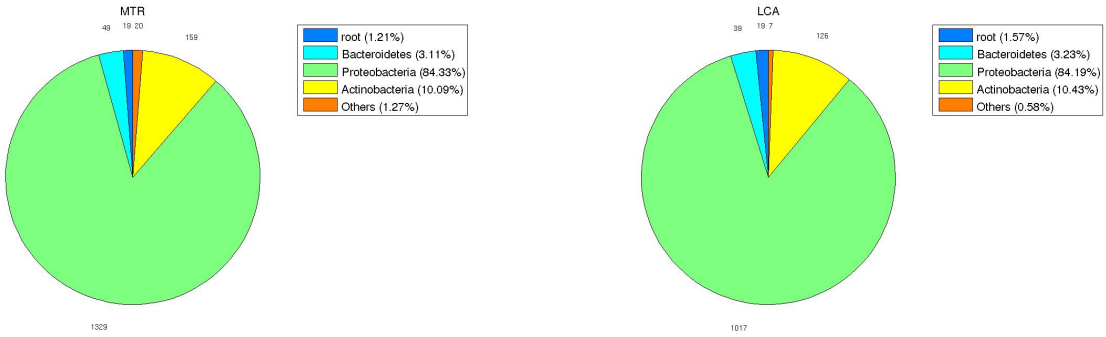

Fig. 19. Population distributions (rank Phylum) of Saltern dataset by MTR (*left*) and LCA (*right*).

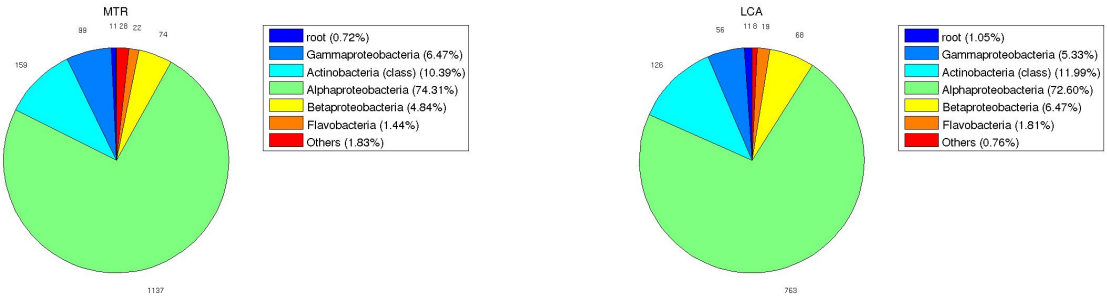

Fig. 20. Population distributions (rank Class) of Saltern dataset by MTR (*left*) and LCA (*right*).

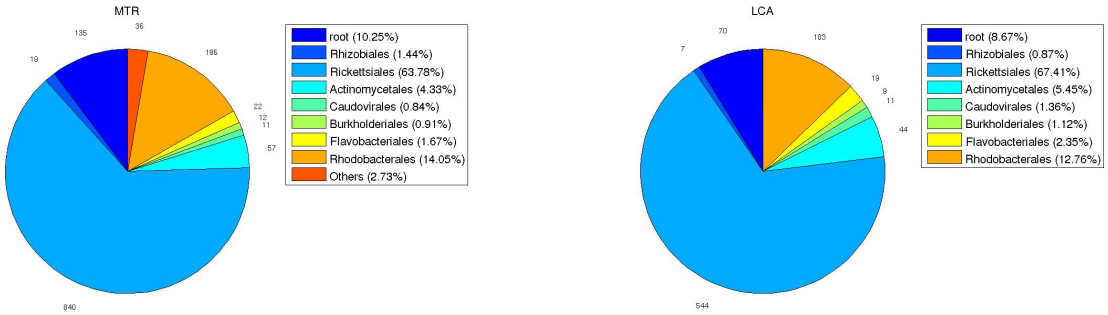

Fig. 21. Population distributions (rank Order) of saltern dataset by MTR (*left*) and LCA (*right*).

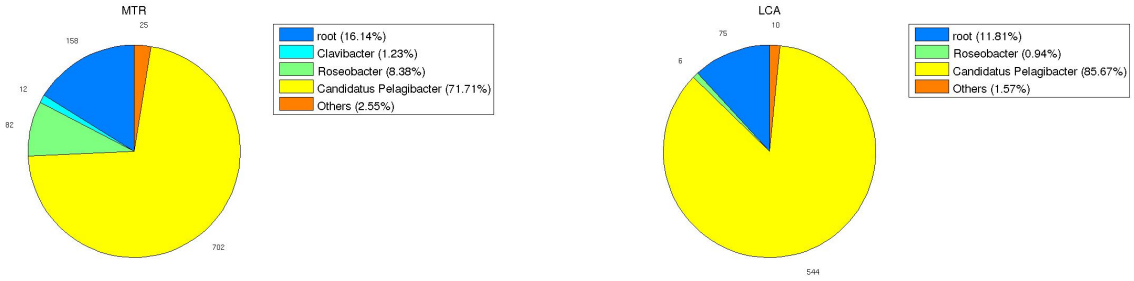

Fig. 22. Population distributions (rank Genus) of Saltern dataset by MTR (*left*) and LCA (*right*).

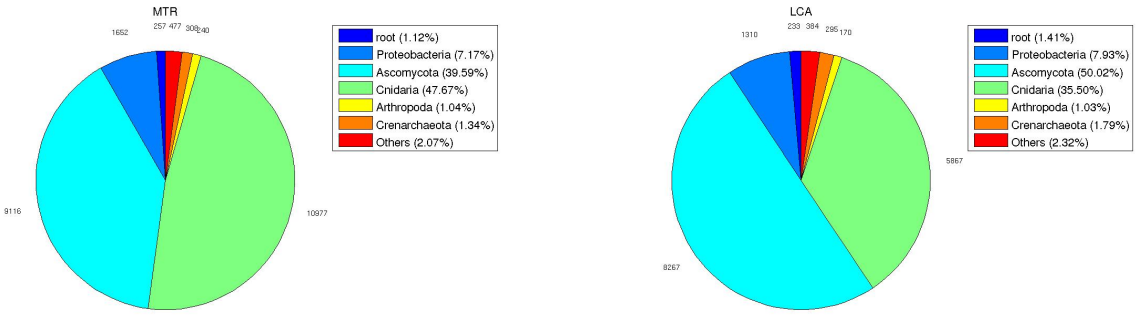

Fig. 23. Population distributions (rank Phylum) of Coral dataset by MTR (*left*) and LCA (*right*).

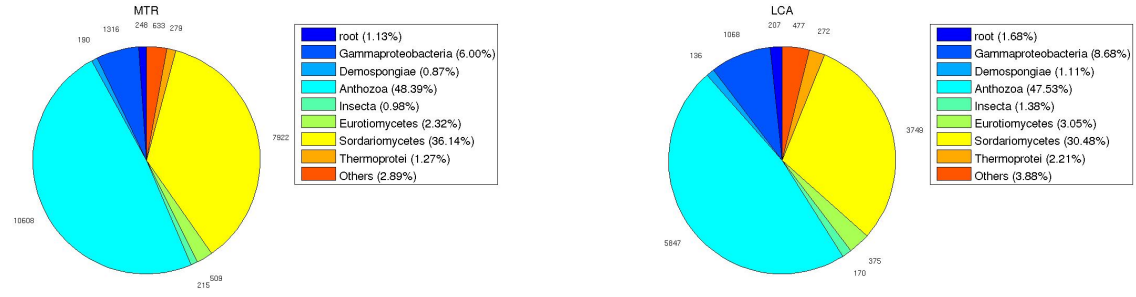

Fig. 24. Population distributions (rank Class) of Coral dataset by MTR (*left*) and LCA (*right*).

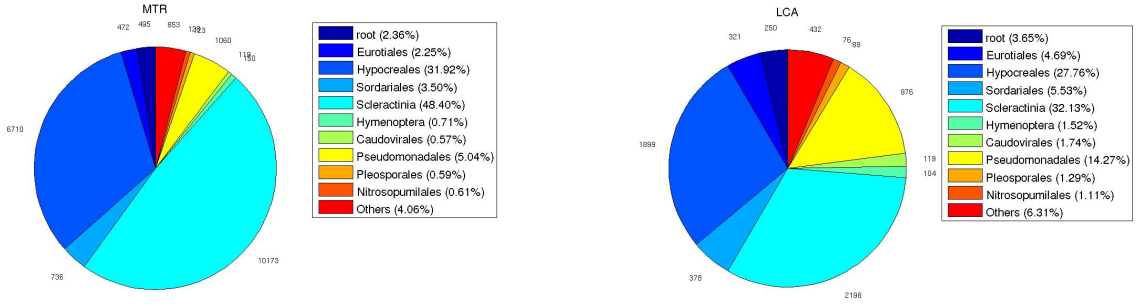

Fig. 25. Population distributions (rank Order) of coral dataset by MTR (*left*) and LCA (*right*).

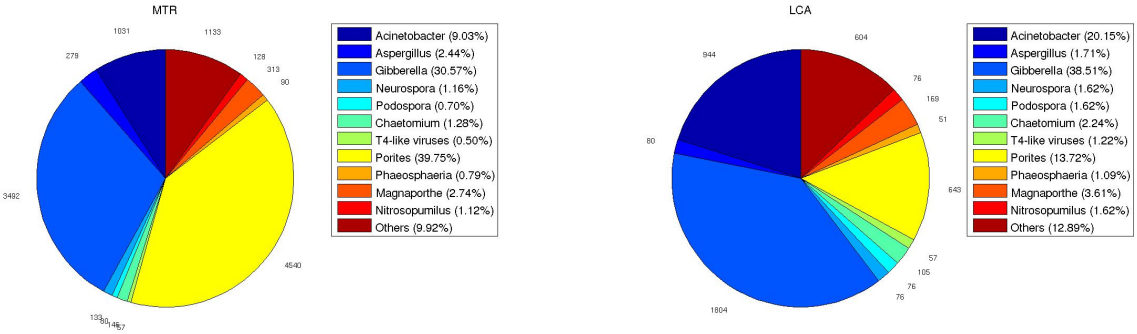

Fig. 26. Population distributions (rank Genus) of Coral dataset by MTR (*left*) and LCA (*right*).

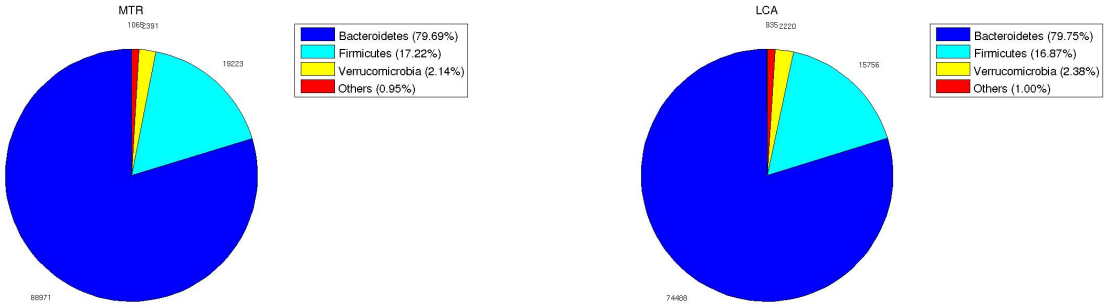

Fig. 27. Population distributions (rank Phylum) of Chicken dataset by MTR (*left*) and LCA (*right*).

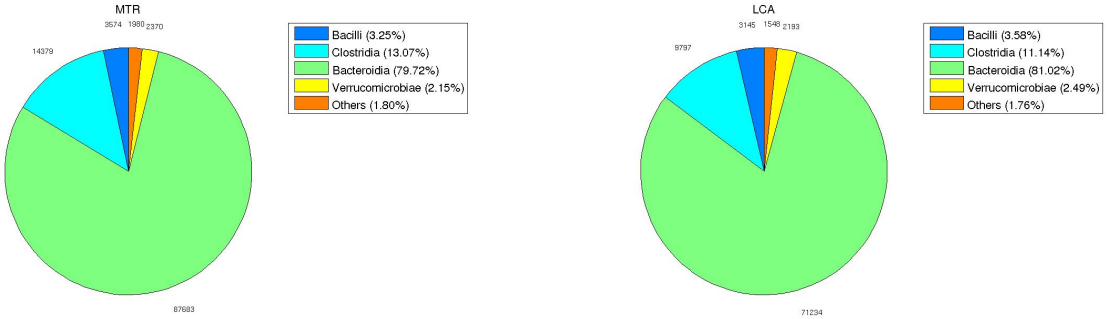

Fig. 28. Population distributions (rank Class) of Chicken dataset by MTR (*left*) and LCA (*right*).

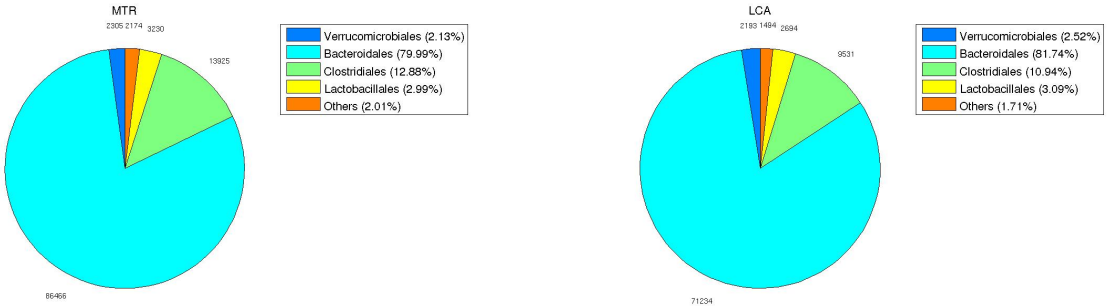

Fig. 29. Population distributions (rank Order) of Chicken dataset by MTR (*left*) and LCA (*right*).

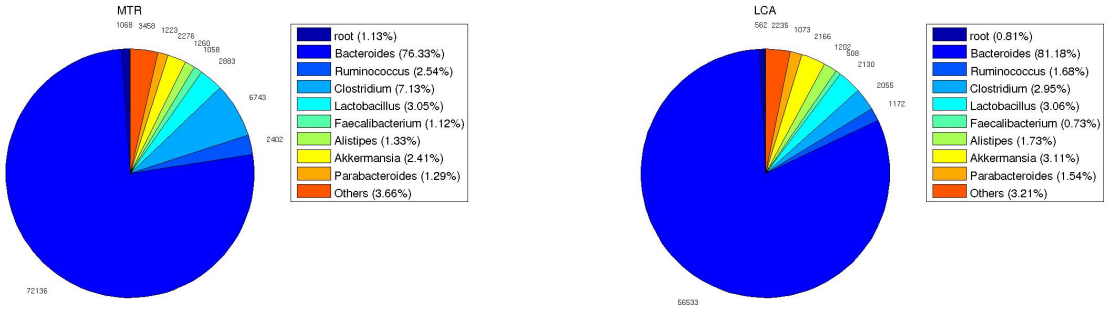

Fig. 30. Population distributions (rank Genus) of Chicken dataset by MTR (*left*) and LCA (*right*).

REFERENCES

Dalevi, D. *et al.* (2008). Annotation of metagenome short reads using proxygenes. *Bioinformatics*, **24**(16).

Edwards, R. *et al.* (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, **7**(1), 57.

Lin, J. (1991). Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Information Theory*, **37**(1), 145–151.

Meyer, F. *et al.* (2008). The metagenomics rast server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**(1), 386.

Rodriguez-Brito, B. *et al.* (2007). Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environmental Microbiology*, **9**(11), 2707–2719.